

CHAPTER |

The Concepts of Power Analysis

The power of a statistical test is the probability that it will yield statistically significant results. Since statistical significance is so earnestly sought and devoutly wished for by behavioral scientists, one would think that the *a priori* probability of its accomplishment would be routinely determined and well understood. Quite surprisingly, this is not the case. Instead, if we take as evidence the research literature, we find that statistical power is only infrequently understood and almost never determined. The immediate reason for this is not hard to discern—the applied statistics textbooks aimed at behavioral scientists, with few exceptions, give it scant attention.

The purpose of this book is to provide a self-contained comprehensive treatment of statistical power analysis from an “applied” viewpoint. The purpose of this chapter is to present the basic conceptual framework of statistical hypothesis testing, giving emphasis to power, followed by the framework within which this book is organized.

1.1 GENERAL INTRODUCTION

When the behavioral scientist has occasion to don the mantle of the applied statistician, the probability is high that it will be for the purpose of testing one or more null hypotheses, i.e., “the hypothesis that the phenomenon to be demonstrated is in fact absent [Fisher, 1949, p. 13].” Not that he hopes to “prove” this hypothesis. On the contrary, he typically hopes to “reject” this hypothesis and thus “prove” that the phenomenon in question is in fact present.

Let us acknowledge at the outset the necessarily probabilistic character of statistical inference, and dispense with the mocking quotation marks

about words like *reject* and *prove*. This may be done by requiring that an investigator set certain appropriate probability standards for research results which provide a basis for rejection of the null hypothesis and hence for the proof of the existence of the phenomenon under test. Results from a random sample drawn from a population will only approximate the characteristics of the population. Therefore, even if the null hypothesis is, in fact, true, a given sample result is not expected to mirror this fact exactly. Before sample data are gathered, therefore, the investigator working in the Fisherian framework selects some prudently small value α (say .01 or .05), so that he may eventually be able to say about his sample data, "If the null hypothesis is true, the probability of the obtained sample result is no more than α ," i.e. a statistically significant result. If he can make this statement, since α is small, he is said to have rejected the null hypothesis "with an α significance criterion" or "at the α significance level." If, on the other hand, he finds the probability to be greater than α , he cannot make the above statement and he has failed to reject the null hypothesis, or, equivalently finds it "tenable," or "accepts" it, all at the α significance level.

We have thus isolated one element of this form of statistical inference, the standard of proof that the phenomenon exists, or, equivalently, the standard of disproof of the null hypothesis that states that the phenomenon does not exist.

Another component of the significance criterion concerns the exact definition of the nature of the phenomenon's existence. This depends on the details of how the phenomenon is manifested and statistically tested, e.g., the directionality/nondirectionality ("one tailed"/"two tailed") of the statement of the alternative to the null hypothesis.¹ When, for example, the investigator is working in a context of comparing some parameter (e.g., mean, proportion, correlation coefficient) for two populations A and B, he can define the existence of the phenomenon in two different ways:

1. The phenomenon is taken to exist if the parameters of A and B differ. No direction of the difference, such as A larger than B, is specified, so that departures in either direction from the null hypothesis constitute evidence against it. Because either tail of the sampling distribution of differences may contribute to α , this is usually called a two-tailed or two-sided test.
2. The phenomenon is taken to exist only if the parameters of A and B differ in a direction specified in advance, e.g., A larger than B. In this

¹ Some statistical tests, particularly those involving comparisons of more than two populations, are naturally nondirectional. In what immediately follows, we consider those tests which contrast two populations, wherein the experimenter ordinarily explicitly chooses between a directional and nondirectional statement of his alternate hypothesis. See below, Chapters 7 and 8.

1.1 GENERAL INTRODUCTION

circumstance, de-
specified constitu-
distribution of di-
tailed or one-side

It is convenient
the probability of
of the definition
the significance c
significance level
(a) that the phen
manifested by any
and (b) that the st
5% of the time if
defining the phen
for A is larger th
rejecting the null
 $\alpha_1 = .10$. The con
into a single ent
combination defin
of the outcome w
the range of valu
gator plans a stat
he has effected a
those which will
risk a no greater t
those which will

The above rev
null hypothesis an
which will lead to
criterion embodie
entire discussion a

But what if, in
false? This is the
null hypothesis fo
that the phenome
exists in the popu

² The author has
tests in psychological
The bases for these
These tests are how
made full provision

quiring that an
for research
sis and hence
results from a
te the charac-
sis is, in fact,
xactly. Before
the Fisherian
(5), so that he
"hypothesis is
than α ," i.e.
nce α is small,
gnificance cri-
l, he finds the
statement and
it "tenable,"

ical inference,
divally, the
phenomenon

ie exact defini-
on the details
sted, e.g., the
of the state-
le, the investi-
r (e.g., mean,
and B, he can

A and B differ.
pecified, so that
stitute evidence
ifferences may
l test.

rs of A and B
an B. In this

more than two
ve consider those
linarily explicitly
nate hypothesis.

circumstance, departures from the null hypothesis only in the direction specified constitute evidence against it. Because only one tail of the sampling distribution of differences may contribute to α , this is usually called a one-tailed or one-sided test.

It is convenient to conceive of the significance criterion as embodying both the probability of falsely rejecting the null hypothesis, α , and the "sidedness" of the definition of the existence of the phenomenon (when relevant). Thus, the significance criterion on a two-tailed test of the null hypothesis at the .05 significance level, which will be symbolized as $\alpha_2 = .05$, says two things: (a) that the phenomenon whose existence is at issue is understood to be manifested by any difference between the two populations' parameter values, and (b) that the standard of proof is a sample result that would occur less than 5% of the time if the null hypothesis is true. Similarly, a prior specification defining the phenomenon under study as that for which the parameter value for A is larger than that of B (i.e., one-tailed) and the probability of falsely rejecting the null is set at .10 would be symbolized as a significance criterion of $\alpha_1 = .10$. The combination of the probability and the sidedness of the test into a single entity, the significance criterion, is convenient because this combination defines in advance the "critical region," i.e., the range of values of the outcome which leads to rejection of the null hypothesis and, perforce, the range of values which leads to its nonrejection. Thus, when an investigator plans a statistical test at some given significance criterion, say $\alpha_1 = .10$, he has effected a specific division of all the possible results of his study into those which will lead him to conclude that the phenomenon exists (with risk α no greater than .10 and a one-sided definition of the phenomenon) and those which will not make possible that conclusion.²

The above review of the logic of classical statistical inference reduces to a null hypothesis and a significance criterion which defines the circumstances which will lead to its rejection or nonrejection. Observe that the significance criterion embodies the risk of mistakenly rejecting a null hypothesis. The entire discussion above is conditional on the truth of the null hypothesis.

But what if, indeed, the phenomenon *does* exist and the null hypothesis is *false*? This is the usual expectation of the investigator, who has stated the null hypothesis for tactical purposes so that he may reject it and conclude that the phenomenon exists. But, of course, the fact that the phenomenon exists in the population far from guarantees a statistically significant result,

² The author has elsewhere expressed serious reservations about the use of directional tests in psychological research in all but relatively limited circumstances (Cohen, 1965). The bases for these reservations would extend to other regions of behavioral science. These tests are however of undoubted statistical validity and in common use, so he has made full provision for them in this work.

i.e., one which warrants the conclusion that it exists, for this conclusion depends upon meeting the agreed-upon standard of proof (i.e., significance criterion). It is at this point that the concept of statistical power must be considered.

The power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists. Given the characteristics of a specific statistical test of the null hypothesis and the state of affairs in the population, the power of the test can be determined. It clearly represents a vital piece of information about a statistical test applied to research data (cf. Cohen, 1962). For example, the discovery, during the planning phase of an investigation, that the power of the eventual statistical test is low should lead to a revision in the plans. As another example, consider a completed experiment which led to nonrejection of the null hypothesis. An analysis which finds that the power was low should lead one to regard the negative results as ambiguous, since failure to reject the null hypothesis cannot have much substantive meaning when, even though the phenomenon exists (to some given degree), the *a priori* probability of rejecting the null hypothesis was low. A detailed consideration of the use of power analysis in planning investigations and assessing completed investigations is reserved for later sections.

The power of a statistical test depends upon three parameters: the significance criterion, the reliability of the sample results, and the "effect size," that is, the *degree* to which the phenomenon exists.

1.2 SIGNIFICANCE CRITERION

The role of this parameter in testing null hypotheses has already been given some consideration. As noted above, the significance criterion represents the standard of proof that the phenomenon exists, or the risk of mistakenly rejecting the null hypothesis. As used here, it directly implies the "critical region of rejection" of the null hypothesis, since it embodies both the probability of a class of results given that the null hypothesis is true (α), as well as the definition of the phenomenon's existence with regard to directionality.

The significance level, α , has been variously called the error of the first kind, the Type I error, and the alpha error. Since it is the rate of rejecting a true null hypothesis, it is taken as a relatively small value. It follows then that the smaller the value, the more rigorous the standard of null hypothesis rejection or, equivalently, of proof of the phenomenon's existence. Assume that a phenomenon exists in the population to some given degree. Other things equal, the more stringent the standard for proof, i.e., the lower the value of α , the poorer the chances are that the sample will provide results

1.2 SIGNIFICANCE

which meet this
gator is prepared
sis, the probabil
be the case were
of false rejection

The practice
results in power
the power ($1 - \beta$)
or beta error, si
null hypothesis.
weighing, in a m
these two kinds
pothesis rejection
he may reduce t
is $1 - .10 = .90$.

1. The gen
science may we
realize that the
(Cohen, 1962).
would lead to a
revision of the a

2. If the in
ception of the re
rejection to risk
he implicitly bel
assumed conditi
In another situa
the relative seri
thus mistaken r
serious as mista

The directio
above examples
hypothesis can l
region is in *both*
a *t* ratio), the r
level which is d
predicted. Since
null hypothesis
no power to de
predicted direct
equal for all pra

, for this conclusion
proof (i.e., significance
statistical power must be

the probability that it
probability that it will
even the characteristics
the state of affairs in
1. It clearly represents
plied to research data
the planning phase of
ical test is low should
consider a completed
pothesis. An analysis
o regard the negative
ypothesis cannot have
henomenon exists (to
he null hypothesis was
sis in planning investi-
ved for later sections.
parameters: the signi-
and the "effect size,"

eses has already been
icance criterion repre-
ts, or the risk of mis-
it directly implies the
since it embodies both
ypothesis is true (α), as
ith regard to direction-

d the error of the first
the rate of rejecting a
ue. It follows then that
ard of null hypothesis
on's existence. Assume
ie given degree. Other
oof, i.e., the lower the
ple will provide results

which meet this standard, i.e., the lower the power. Concretely, if an investi-
gator is prepared to run only a 1% risk of false rejection of the null hypothe-
sis, the probability of his data meeting this standard is lower than would
be the case were he prepared to use the less stringent standard of a 10% risk
of false rejection.

The practice of taking a very small ("the smaller the better") then
results in power values being relatively small. However, the complement of
the power ($1 - \text{power}$), here symbolized as b , is also error, called Type II
or beta error, since it represents the "error" rate of failing to reject a false
null hypothesis. Thus it is seen that statistical inference can be viewed as
weighing, in a manner relevant to the substantive issues of an investigation,
these two kinds of errors. An investigator can set the risk of false null hy-
pothesis rejection at a vanishingly small level, say $\alpha = .001$, but in so doing,
he may reduce the power of his test to .10 (hence beta error probability, b ,
is $1 - .10 = .90$). Two comments may be made here:

1. The general neglect of issues of statistical power in behavioral
science may well result, in such instances, in the investigator's failing to
realize that the $\alpha = .001$ value leads in his situation to $\text{power} = .10$, $b = .90$
(Cohen, 1962). Presumably, although not necessarily, such a realization
would lead to a revision of experimental plans, including possibly an upward
revision of the α level to increase power.

2. If the investigator proceeds as originally planned, he implies a con-
ception of the relative seriousness of Type I to Type II error (risk of false null
rejection to risk of false null acceptance) of $b/\alpha = .90/.001 = 900$ to 1, i.e.,
he implicitly believes that mistakenly rejecting the null hypothesis under the
assumed conditions is 900 times more serious than mistakenly accepting it.
In another situation, with $\alpha = .05$, $\text{power} = .80$, and hence $b = 1 - .80 = .20$,
the relative seriousness of Type I to Type II error is $b/\alpha = .20/.05 = 4$ to 1;
thus mistaken rejection of the null hypothesis is considered four times as
serious as mistaken acceptance.

The directionality of the significance criterion (left unspecified in the
above examples) also bears on the power of a statistical test. When the null
hypothesis can be rejected in *either* direction so that the critical significance
region is in *both* tails of the sampling distribution of the test statistic (e.g.,
a t ratio), the resulting test will have less power than a test at the same α
level which is directional, *provided that* the sample result is in the direction
predicted. Since directional tests cannot, by definition, lead to rejecting the
null hypothesis in the direction *opposite* to that predicted, these tests have
no power to detect such effects. When the experimental results are in the
predicted direction, all other things equal, a test at level α_1 will have power
equal for all practical purposes to a test at $2\alpha_2$.

Concretely, if an experiment is performed to detect a difference between the means of populations A and B, say m_A and m_B , in *either* direction at the $\alpha_2 = .05$ significance criterion, under given conditions, the test will have a certain power. If, instead, an anticipation of m_A greater than m_B leads to a test at $\alpha_1 = .05$, this test will have power approximately equal to a two-tailed test with $\alpha_2 = .10$, hence greater power than the test at $\alpha_2 = .05$, provided that in fact m_A is greater than m_B . If m_B is greater than m_A , the test at $\alpha_1 = .05$ has *no* power, since that conclusion is inadmissible. The temptation to perform directional tests because of their greater power at the same α level should be tempered by the realization that they preclude finding results opposite to those anticipated. There are occasional circumstances where the nature of the decision is such that the investigator does not need to know about effects in the opposite direction. For example, he will take a certain course of action if m_A is greater than m_B and not otherwise. If otherwise, he does not need to distinguish between their equality and m_B greater than m_A . In such infrequent instances, one-tailed tests are appropriate (Cohen, 1965, pp. 106-111).

In the tables in this book, provision is made for tests at the .01, .05, and .10 significance levels. Where a statistical test may ordinarily be performed either nondirectionally or directionally, both α_2 and α_1 tables are provided. Since power for $\alpha_1 = .05$ is virtually identical with power for $\alpha_2 = .10$, a single power table suffices. Similarly, tables for $\alpha_1 = .01$ provide values for $\alpha_2 = .02$, and tables for $\alpha_1 = .10$ values for $\alpha_2 = .20$; also, tables for $\alpha_2 = .01$ provide values for $\alpha_1 = .005$, tables at $\alpha_2 = .05$ provide values for $\alpha_1 = .025$.

1.3 RELIABILITY OF SAMPLE RESULTS AND SAMPLE SIZE

The reliability (or precision) of a sample value is the closeness with which it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown. Depending upon the statistic in question, and the specific statistical model on which the test is based, reliability may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is *always* dependent upon the size of the sample.

For example, one conventional means for assessing the reliability of a statistic is the standard error (SE) of the statistic. If we consider the arithmetic mean of a variable X (\bar{X}), its reliability may be estimated by the standard error of the mean,

$$SE_{\bar{X}} = \sqrt{\frac{s^2}{n}}$$

where s^2 is the usual unbiased estimate (from the random sample) of the

1.3 RELIABILITY OF

population variance (the size of) the sample

Concretely, if a sample of 196, then the standard error of the mean is

Thus, sample mean as measured by the degree to which less any of them of the sample in unit of measurement mean or (to be precise) the coefficient of correlation

On the other hand, the coefficient of correlation

where

r_p = the population correlation coefficient
 n = the number of samples

Note that the reliability of the (generally unbiased) estimate of the correlation coefficient which the correlation coefficient

Not all statistics of a sample value have the same reliability. More precisely, the reliability may be dependent upon the nature of the statistic

The nature of the statistic is illustrated by the following examples: (1) the precision of a sample value is intuitively evident when the sample value is being equal, the sample size is large, i.e., the more the background information is directly formulated intuitively obvious probability of error

Focusing on the sample value

ference between direction at the test will have a m_B leads to a 1 to a two-tailed .05, provided that test at $\alpha_1 = .05$ ation to perform a level should be opposite to those e nature of the about effects in ourse of action if does not need to n such infrequent . 106-111).

t the .01, .05, and rely be performed bles are provided. er for $\alpha_2 = .10$, a provide values for tables for $\alpha_2 = .01$ lues for $\alpha_1 = .025$.

the closeness with opulation value. It opulation value is question, and the lity may or may not opulation value, and s *always* dependent

the reliability of a onsider the arithmet- ated by the standard

idom sample) of the

population variance of X , and n is the number of independent units in (i.e., the size of) the sample.

Concretely, if a sample of $n = 49$ cases yields a variance estimate for IQ of 196, then the standard error of the mean is given by

$$SE_{\bar{x}} = \sqrt{\frac{s^2}{n}} = \sqrt{\frac{196}{49}} = 2.$$

Thus, sample means based on 49 cases can be expected to have variability as measured by their own standard deviation of 2 IQ units. Clearly the greater the degree to which means of different samples vary among themselves, the less any of them can be relied upon, i.e., the less the reliability of the mean of the sample in hand. Note that in this instance reliability depends upon the unit of measurement (IQ) and sample size, but not on the value of the population mean or (to any material degree) on the shape of the IQ distribution.

On the other hand, consider the sampling reliability of a product moment coefficient of correlation, r . Its standard error is

$$SE_r = \frac{1 - r_p^2}{\sqrt{n - 1}},$$

where

r_p = the population value of r , and

n = the number of paired observations in the sample.

Note that the reliability of the sample r depends upon the magnitude of the (generally unknown) population r_p value and n , but not on the units in which the correlated variables are measured.

Not all statistical tests involve the explicit definition of a standard error of a sample value, but all do involve the more general conception of sample reliability. Moreover, and most important, whatever else sample reliability may be dependent upon, it *always* depends upon the size of the sample.

The nature of the dependence of reliability upon n is obvious from the illustrative formulas, and, indeed, intuitively. The larger the sample size, other things being equal, the smaller the error and the greater the reliability or precision of the results. The further relationship with power is also intuitively evident: the greater the precision of the sample results, other things being equal, the greater the probability of detecting a nonnull state of affairs, i.e., the more clearly the phenomenon under test can manifest itself against the background of (experimentally irrelevant) variability. Thus, we can directly formulate the relationship between sample size and power. As is intuitively obvious, increases in sample size increase statistical power, the probability of detecting the phenomenon under test.

Focusing on sample size as an invariant factor in power should not make

✱

1.4 THE EFFECT SIZE

³ The assumption is made here that .50 is the proportion of males in the population of interest.

⁴ For the sake of simplicity, the null hypothesis is treated in this section for the non-directional form of the significance criterion. For example, a directional (one-tailed) test here that the male proportion is greater than .50 implies a null hypothesis that it is equal to or less than .50. The reader may supply his own necessary qualifications of the null hypothesis for the directional case in each illustration.

the diagnosis is zero. Born in multiple population in qu (mean), again a s multiple birth on in a study of the version-extroverts measure for a sam here is that the p other is zero.

Statistical tests that imply the conditional statement: If H_0 is true, then T is small. For example, the t -test for a mean has as its null hypothesis that the population mean is zero, and so, early, a test of whether a mean is zero can be performed by looking at the t -value. The condition is that the variance is known, a condition which is rarely met in practice. In instances we can test for differences in the means (have an effect size) of a variable.

Thus, we see by a null hypothesis meter, one which manifested. With convenient to use the phenomenon

ts potentially
or, be it due
testtubes, or
observations
sion of sample
reduces the
iability which
y will serve to
holly devoted
se of precision
of null hypoth-

power analyses
zed experimen-
ch as the effects
e not explicitly
e precision, is,
the major kinds
le size is one of
the sample size
desired level of
ice criterion and

er statistical test
ent (null hypoth-
specific value for
ine whether there
, the investigator
he relevant popu-
thesis being tested
value.^{3,4} Equiva-
on the presence of

les in the population

s section for the non-
ional (one-tailed) test
othesis that it is equal
ifications of the null

the diagnosis is zero. In another study concerned with the IQs of children born in multiple births, the null hypothesis might be that the multiple birth population in question has a mean IQ of 100 (i.e., the general population mean), again a specific value, or that the size of effect of being part of a multiple birth on IQ is zero. As yet another example of a one-sample test, in a study of the construct validity of a neurophysiological measure of introversion-extroversion, its product moment r with an accepted questionnaire measure for a sample of college students is determined. The null hypothesis here is that the population r is zero, or that the effect size of either on the other is zero.

In circumstances where two populations are being compared, the null hypothesis usually takes the form "the difference in the value of the relevant parameters is zero," a specific value. Thus, in a consumer survey research to determine whether preference for a particular brand A over its chief competitor B is related to the income level of the consumer, the null hypothesis might be: The difference in median family income of brand A and brand B users is zero, or, equivalently, that the size of the effect of income on brand preference is zero. Or, in a personnel selection study to determine which of two screening tests, A or B, is a better predictor of performance ratings (C), the null hypothesis might take the form: The difference between population product moment r 's of A with C and B with C is zero.

Statistical tests involving more than two samples test null hypotheses that imply the constancy of a parameter over the populations involved. The literal statement of the null hypothesis depends upon the specific test involved. For example, the F test of the analysis of variance for $k \geq 2$ means has as its null hypothesis the proposition that the variance of a set of population means is zero, a condition that can only obtain when they are equal. Similarly, a test of whether a set of $k \geq 2$ population proportions are equal can be performed by means of the chi-square statistic. The null hypothesis here is that the variance of the population proportions equals zero (an exact value), a condition which can only obtain when they are all equal. In both of these instances we can think of the null hypothesis as the circumstance in which differences in the independent variable, the k populations, have no effect (have an effect size of zero) on the means or proportions of the dependent variable.

Thus, we see that the absence of the phenomenon under study is expressed by a null hypothesis which specifies an exact value for a population parameter, one which is appropriate to the way the phenomenon under study is manifested. Without intending any necessary implication of causality, it is convenient to use the phrase "effect size" to mean "the *degree* to which the phenomenon is present in the population," or "the degree to which the

null hypothesis is false." Whatever the manner of representation of a phenomenon in a particular research in the present treatment, the null hypothesis always means that the effect size is zero.

By the above route, it can now readily be made clear that when the null hypothesis is false, it is false to some specific degree, i.e., *the effect size (ES) is some specific nonzero value in the population*. The larger this value, the greater the *degree* to which the phenomenon under study is manifested. Thus, in terms of the previous illustrations:

1. If the percentage of males in the population of psychiatric patients bearing a diagnosis of paranoid schizophrenia is 52%, and the effect is measured as a departure from the hypothesized 50%, the ES is 2%; if it is 60%, the ES is 10%, a larger ES.
2. If children of multiple births have a population mean IQ of 96, the ES is 4 IQ units (or -4, depending on directionality of significance criterion); if it is 92, the ES is 8 (or -8) IQ units, i.e., a larger ES.
3. If the population product moment r between neurophysiological and questionnaire measures of introversion-extroversion is .30, the ES is .30; if the r is .60, so is the ES, a larger value and a larger departure from the null hypothesis, which here is $r = 0$.
4. If the population of consumers preferring brand A has a median annual income \$700 higher than that of brand B, the ES is \$700. If the population median difference and hence the ES is \$1000, the effect of income on brand preference would be larger.

Thus, whether measured in one unit or another, whether expressed as a difference between two population parameters or the departure of a population parameter from a constant or in any other suitable way, the ES can itself be treated as a parameter which takes the value zero when the null hypothesis is true and *some other specific nonzero value* when the null hypothesis is false, and in this way the ES serves as an index of degree of departure from the null hypothesis.

The reasons that the above discussion has proceeded in such redundant detail are twofold. On the one hand, ES is in practice a most important determinant of power or required sample size or both, and on the other hand, it is the least familiar of the concepts surrounding statistical inference among practicing behavior scientists. The reason for the latter, in turn, can be found in the difference in null hypothesis testing between the procedures of Fisher (1949) and those of Neyman and Pearson (1928, 1933).

The Fisherian formulation posits the null hypothesis as described above, i.e., the ES is zero, to which the "alternative" hypothesis is that the ES is *not* zero, i.e., *any* nonzero value. Without further specification, although null hypotheses may be tested and thereupon either rejected or not rejected,

no basi
Pearson
size of t
tive hyp
in statist

Thus
lation ES
physiolo
.30, the l

The n
The large
being equ
between
things (si
sample si

To thi
which car
given stat
appropria
illustratio
a departur
two media
units as fa
tists. From
various res
the ideal. A
the way, ev
in terms so
defeating.

Howeve
pare a set c
works. That
must use the
mean weigh
performed. t
are also, for
Thus, as will
lation means
standard dev
"raw" score
(1973), or the
test for $k \geq 2$

no basis for statistical power analysis exists. By contrast, the Neyman-Pearson formulation posits an *exact* alternative for the ES, i.e., the *exact* size of the effect the experiment is designed to detect. With an exact alternative hypothesis or specific nonzero ES to be detected, given the other elements in statistical inference, statistical power analysis may proceed.

Thus, in the previous illustrations, the statements about possible population ES values (e.g., "if the population product moment r between neurophysiological and questionnaire measures of introversion-extroversion is .30, the ES is .30") are statements of alternative hypotheses.

The relationship between ES and power should also be intuitively evident. The larger the ES posited, other things (significance criterion, sample size) being equal, the greater the power of the test. Similarly, the relationship between ES and necessary sample size: the larger the ES posited, other things (significance criterion, desired power) being equal, the smaller the sample size necessary to detect it.

To this point, the ES has been considered quite abstractly as a parameter which can take on varying values (including zero in the null case). In any given statistical test, it must be indexed or measured in some defined unit appropriate to the data, test, and statistical model employed. In the previous illustrations, ES was variously expressed as a departure in percent from 50, a departure in IQ units from 100, a product moment r , a difference between two medians in dollars, etc. It is clearly desirable to reduce this diversity of units as far as possible, consistent with present usage by behavioural scientists. From one point of view, a universal ES index, applicable to all the various research issues and statistical models used in their appraisal, would be the ideal. Apart from some formidable mathematical-statistical problems in the way, even if such an ideal could be achieved, the result would express ES in terms so unfamiliar to the researcher in behavioral science as to be self-defeating.

However, some generalization is obviously necessary. One cannot prepare a set of power tables for each new measurement unit with which one works. That is, the researcher who plans a test for a difference in mean IQs must use the same power tables as another who plans a test for a difference in mean weights, just as they will use the same tables of t when the research is performed. t is a "pure" (dimensionless) number, one free of raw unit, as are also, for example, correlation coefficients or proportions of variance. Thus, as will be seen in Chapter 2, the ES index for differences between population means is standardized by division by the common within-population standard deviation (σ), i.e., the ES here is not the difference between mean "raw" scores, but the difference between mean "z" standard scores (Hays, 1973), or the mean difference expressed in within-population σ units. In the F test for $k \geq 2$ population means, the ES also uses such standardized means;

in testing "main effects" in the analysis of variance the ES is *their* standard deviation, σ_m , the standard deviation of standardized means (Chapter 8).

Each test for which power tables are provided thus has a metric-free ES index appropriate to it. A higher order of generalization is frequently possible. Specifically, several ES indices can be translated into the proportion of variance (PV) accounted for in the dependent variable. Where this is possible, it is discussed in the introductory material for the test. Also, each ES index chosen usually relates to yet other commonly used indices and these are also described in the same place.

The behavior scientist who comes to statistical power analysis may find himself grappling with the problem of what ES to posit as an alternate to the null hypothesis, or, more simply, how to answer the questions "How large an effect do I expect exists in the population?" He may initially find it difficult to answer the question even in general terms, i.e., "small" or "large," let alone in terms of the specific ES index demanded. Being forced to think in more exact terms than demanded by the Fisherian alternative (ES is any nonzero value) is likely to prove salutary. He can call upon theory for some help in answering the question and on his critical assessment of prior research in the area for further help. When these are supplemented with the understanding of the ES index provided in the introductory material to the relevant chapter, he can decide upon the ES value to adopt as an alternative to the null.

When the above has not provided sufficient guidance, the reader has an additional recourse. For each statistical test's ES index, the author proposes, *as a convention*, ES values to serve as operational definitions of the qualitative adjectives "small," "medium," and "large." This is an operation fraught with many dangers: The definitions are arbitrary, such qualitative concepts as "large" are sometimes understood as absolute, sometimes as relative; and thus they run a risk of being misunderstood.

In justification, several arguments may be offered. It must first be said that all conventions are arbitrary. One can only demand of them that they not be unreasonable. Also, all conventions may be misused and their conventional status thus abused. For example, the .05 significance criterion, although unofficial, has come to serve as a convention for a (minimum) basis for rejecting the null hypothesis in most areas of behavioral and biological science. Unfortunately, its status as only a convention is frequently ignored; there are many published instances where a researcher, in an effort at rectitude, fails to report that a much desired null rejection would be possible at the .06 level but instead treats the problem no differently than he would have had it been at the .50 level! Still, it is convenient that "significance" without further specification can be taken to mean "significance at no more than the .05 level."

Although able by reas criteria to u sizes such as not be so sm ment and e large as to Many effect are likely to in validity o involved. In quest by st Tukey's deli size between approach vi to encroach frequently a physiological or the prese

Since eff the control research de: simple exan Assume tha population A research randomized operating w comparing means). No ence of 4 sc families affc the brother- will be redi siblings = .4 a larger val sizes may b mental tech

Each o appropriate into alterna "large" wi

their standard
is (Chapter 8).
metric-free ES
frequently pos-
e proportion of
ere this is pos-
Also, each ES
es and these are

alysis may find
an alternate to
questions "How
ay initially find
e., "small" or
d. Being forced
erian alternative
call upon theory
al assessment of
plemented with
tory material to
pt as an alterna-

ne reader has an
author proposes,
of the qualitative
ion fraught with
tive concepts as
as relative; and

first be said that
m that they not
d their conven-
terion, although
) basis for reject-
ological science.
y ignored; there
fort at rectitude,
ossible at the .06
ould have had it
" without further
ore than the .05

Although arbitrary, the proposed conventions will be found to be reasonable by reasonable people. An effort was made in selecting these operational criteria to use levels of ES which accord with a subjective average of effect sizes such as are encountered in behavioral science. "Small" effect sizes must not be so small that seeking them amidst the inevitable operation of measurement and experimental bias and lack of fidelity is a bootless task, yet not so large as to make them fairly perceptible to the naked observational eye. Many effects sought in personality, social, and clinical-psychological research are likely to be small effects as here defined, both because of the attenuation in validity of the measures employed and the subtlety of the issues frequently involved. In contrast, large effects must not be defined as so large that their quest by statistical methods is wholly a labor of supererogation, or to use Tukey's delightful term "statistical sanctification." That is, the difference in size between apples and pineapples is of an order which hardly requires an approach via statistical analysis. On the other side, it cannot be defined so as to encroach on a reasonable range of values called medium. Large effects are frequently at issue in such fields as sociology, economics, and experimental and physiological psychology, fields characterized by the study of potent variables or the presence of good experimental control or both.

Since effects are appraised against a background of random variation, the control of various sources of variation through the use of improved research designs serves to increase effect sizes as they are defined here. A simple example of this is a study of sex difference in some defined ability. Assume that a difference of 4 score points exists between male and female population means, where each population has a standard deviation of 16. A research plan which randomly samples the two populations (simple randomized design or comparison between two independent means) is operating with an ES of $4/16 = .25$. Another research plan might proceed by comparing means of males and their sisters (comparison of two dependent means). Now, these populations can also be assumed to have a mean difference of 4 score points, but because of the removal of the variation between families afforded by this design (or equivalently when allowance is made for the brother-sister correlation in the ability), the *effective* standard deviation will be reduced to the fraction $\sqrt{1-r}$ of 16, say to 12 (when r between siblings = .44), and the actual ES operating in the situation is $4/12 = .33$, a larger value than for the simple randomized design. Thus, *operative* effect sizes may be increased not only by improvement in measurement and experimental technique, but also by improved experimental designs.

Each of the Chapters 2-8 will present in some detail the ES index appropriate to the test to which the chapter is devoted. Each will be translated into alternative forms, the operational definitions of "small," "medium," and "large" will be presented, and examples drawn from various fields will

illustrate the test. This should serve to clarify the ES index involved and make the methods and tables useful in research planning and appraisal.

1.5 TYPES OF POWER ANALYSIS

Four parameters of statistical inference have been described: power, significance criterion (α), sample size (n), and effect size (ES). They are so related that any one of them is a function of the other three, which means that when any three of them are fixed, the fourth is completely determined. This relationship makes formally possible four types of power analysis; in each, one of these parameters is determined as a function of the other three (Cohen, 1965, pp. 97-101).

1.5.1 POWER AS A FUNCTION OF α , ES, AND n . The preceding material has been largely oriented toward the type of analysis in which, given the specification of α , ES, and n , power is determined. For example, an investigator plans a test of the significance of a product moment r at $\alpha_2 = .05$ using $n = 30$ cases. The ES he wishes to detect is a population r of .40. Given these specifications, he finds (by the methods of Section 3.3 in Chapter 3) that power equals .61. He may then decide to change his specifications to increase power.

Such analyses are usefully performed as part of research planning. They can also be performed on completed studies to determine the power which a given statistical test had, as in the power survey of the studies in a volume of the *Journal of Abnormal and Social Psychology* (Cohen, 1962). In each of Chapters 2-9, the power tables (numbered B.3.A, where B is the chapter number and A indexes the significance criterion) are designed for this type of analysis. The sections designated B.3 discuss and illustrate the use of these tables.

1.5.2 n AS A FUNCTION OF ES, α , AND POWER. When an investigator anticipates a certain ES, sets a significance criterion α , and then specifies the amount of power he desires, the n which is necessary to meet these specifications can be determined. This (second) type of power analysis must be at the core of any rational basis for deciding on the sample size to be used in an investigation (Cohen, 1965, pp. 97-99). For example, an investigator wishes to have power equal to .80 to detect a population r of .40 (the ES) at $\alpha_2 = .05$. By the methods described in Section 3.4 in Chapter 3, he finds that he must have $n = 46$ cases to meet these specifications. (A discussion of the basis for specifying desired power and the use of power = .80 as a convention will be found in Section 2.4 of Chapter 2.)

This major type of power analysis is discussed and illustrated in the Sections B.4 (where B indexes the chapter numbers 2-8). Each of these sections contain sample size tables (numbered B.4.A) from which, given α ,

1.5 TYPES OF

the ES, and de
to n determin

1.5.3 ES
analysis is of
quite useful in
which one can
example, an in
product mom
tion r (the ES
specifications
3.3.5) is that t
46, the detect

This form
sons of resear
can define, as
that ES detect
test. So defin
test, expressed

This type
chapters. Ho
the tables, he
tests discusse
proves more

1.5.4 α
analysis answ
given ES wit
an investigat
power to be
specification
be found (b
about $\alpha_1 = .$

This typ
strength of
loath to cor
means toler
power. Wh
stances may

This typ
2-9, althoug
reader has

dex involved and
nd appraisal.

described: power,
(ES). They are so
free, which means
pletely determined.
power analysis; in
of the other three

preceding material
which, given the
sample, an investi-
r at $\alpha_2 = .05$ using
of .40. Given these
apter 3) that power
to increase power.
research planning.
etermine the power
y of the studies in
ogy (Cohen, 1962).
3.A, where B is the
n) are designed for
s and illustrate the

hen an investigator
and then specifies
sary to meet these
power analysis must
e sample size to be
example, an investi-
ulation r of .40 (the
.4 in Chapter 3, he
ations. (A discussion
of power = .80 as a

d illustrated in the
2-8). Each of these
rom which, given α ,

the ES, and desired power, the n is determined. A slightly different approach to n determination is employed in Chapter 9.

1.5.3 ES AS A FUNCTION OF α , n , AND POWER. A third type of power analysis is of less general utility than the first two, but may nevertheless be quite useful in special circumstances (Cohen, 1970). Here, one finds the ES which one can expect to detect for given α , n , and with specified power. For example, an investigator may pose the question, "For a significance test of a product moment r at $\alpha_2 = .05$ with a sample of $n = 30$, what must the population r (the ES) be if power is to be .80, i.e., what is the *detectable* ES for these specifications?" The answer, obtainable by backward interpolation (in Table 3.3.5) is that the population r must be approximately .48. Were his n equal to 46, the detectable ES would be $r = .40$.

This form of power analysis may be conventionalized for use in comparisons of research results as in literature surveys (Cohen, 1965, p. 100). One can define, as a convention, a comparative detectable effect size (CDES) as that ES detectable at $\alpha_2 = .05$ with power = .50 for the n used in the statistical test. So defined, the CDES is an inverse measure of the sensitivity of the test, expressed in the appropriate ES unit.

This type of power analysis is not discussed in detail in the ensuing chapters. However, when the reader has become familiar with the use of the tables, he will find that it can be accomplished for all of the statistical tests discussed by backward interpolation in the power tables, or when it proves more convenient, in the sample size tables.

1.5.4 α AS A FUNCTION OF n , POWER, AND ES. The last type of power analysis answers the question, "What significance level must I use to detect a given ES with specified probability (power) for a fixed given n ?" Consider an investigator whose anticipated ES is a population r of .30, who wishes power to be .75, and who has an n of 50, which he cannot increase. These specifications determine the significance criterion he must use, which can be found (by rough interpolation between subtables in Table 3.4.1) to be about $\alpha_1 = .08$, or $\alpha_2 = .15$.

This type of analysis is very uncommon, at least partly because of the strength of the significance criterion convention, which makes investigators loath to consider "large" values of α . We have seen that this frequently means tolerating (usually without knowing it) large values of b , i.e., low power. When power issues are brought into consideration, some circumstances may dictate unconventionally large α criteria (Cohen, 1965, p. 99ff).

This type of power analysis is not, as such, further discussed in Chapters 2-9, although it is indirectly considered in some of the examples. When the reader has become familiar with the tables, it can be accomplished for all

the statistical tests discussed in this book by interpolation between subtables of the sample size tables (B.4.A), or when more convenient, between power tables (B.3.A), within the range provided for α , i.e., α_2 : .01-.20, and α_1 : .005-.10.

In summary, four types of power analysis have been described. This book is designed primarily to facilitate two of these, the solutions for power and for sample size. It is also possible, but with less ease, to accomplish the other two, solution for ES and for α , by means of backward interpolation in the tables.

1.5.5 "PROVING" THE NULL HYPOTHESIS. Research reports in the literature are frequently flawed by conclusions that state or imply that the null hypothesis is true. For example, following the finding that the difference between two sample means is not statistically significant, instead of properly concluding from this failure to reject the null hypothesis that the data do not warrant the conclusion that the population means differ, the writer concludes, at least implicitly, that there is *no* difference. The latter conclusion is always strictly invalid, and is functionally invalid as well unless power is high. The high frequency of occurrence of this invalid interpretation can be laid squarely at the doorstep of the general neglect of attention to statistical power in the training of behavioral scientists.

What is really intended by the invalid affirmation of a null hypothesis is not that the population ES is literally zero, but rather that it is negligible, or trivial. This proposition may be validly asserted under certain circumstances. Consider the following: for a given hypothesis test, one defines a numerical value i (or *iota*) for the ES, where i is so small that it is appropriate in the context to consider it negligible (trivial, inconsequential). Power ($1 - b$) is then set at a high value, so that b is relatively small. When, additionally, α is specified, n can be found. Now, if the research is performed with this n and it results in nonsignificance, it is proper to conclude that the population ES is no more than i , i.e., that it is negligible; this conclusion can be offered as significant at the b level specified. In much research, "no" effect (difference, correlation) functionally means one that is negligible; "proof" by statistical induction is probabilistic. Thus, in using the same logic as that with which we reject the null hypothesis with risk equal to α , the null hypothesis can be accepted in preference to that which holds that $ES = i$ with risk equal to b . Since i is negligible, the conclusion that the population ES is not as large as i is equivalent to concluding that there is "no" (nontrivial) effect. This comes fairly close and is functionally equivalent to affirming the null hypothesis with a controlled error rate (b), which, as noted above, is what is actually intended when null hypotheses are incorrectly affirmed (Cohen, 1965, pp. 100-101; Cohen, 1970). (See Illustrative Examples 2.9, 3.5, 6.8, and 9.24.)

1.7 P

This
"negati
If, for e
and plai
to detec
the requ
For the
for pow
 $n = 258$
3.4.1). T
bility of
it takes t
no (nont

1.6 SIG

Althc
tionship
computa
which us
the effect
we define
exceeds a
criterion
criterion
the symb
 d_e for the

1.7 PLA

Each c
are simila

Section

Section

Section
their use

Section
of their us

Section
and illustr

interpolation between subtables is convenient, between power α , i.e., α_2 : .01-.20, and α_1 :

have been described. This book provides the solutions for power and α , to accomplish the other backward interpolation in the

Research reports in the that state or imply that the finding that the difference significant, instead of properly hypothesis that the data do on means differ, the writer inference. The latter conclusion valid as well unless power is invalid interpretation can be effect of attention to statistical

tion of a null hypothesis is not whether that it is negligible, or under certain circumstances. test, one defines a numerical that it is appropriate in the sequential). Power $(1 - \beta)$ is small. When, additionally, α is performed with this n and it is deduced that the population ES is concluded that the conclusion can be offered as either, "no" effect (difference, negligible; "proof" by statistical logic as that with which we conclude that the null hypothesis can be rejected if $ES = i$ with risk equal to β . If the population ES is not as large as i (nontrivial) effect. This comes from confirming the null hypothesis as described above, is what is actually affirmed (Cohen, 1965, pp. 2.9, 3.5, 6.8, and 9.24.)

This statistically valid basis for extracting positive conclusions from "negative findings" may not be of much practical help to most investigators. If, for example, one considers a population $r = .10$ as negligible (hence, i), and plans a test of the null hypothesis (at $\alpha_2 = .05$) for power = .95 ($\beta = .05$) to detect i , one discovers that the n required is 1308; for power = .90 ($\beta = .10$), the required $n = 1046$; and for power = .80 ($\beta = .20$), $n = 783$ (Table 3.4.1). For the much more liberal specification of $r = .20$ as i , the test (at $\alpha_2 = .05$) for power = .95 ($\beta = .05$) requires $n = 322$; for power = .90 ($\beta = .10$) requires $n = 258$, and even for power = .80 ($\beta = .20$), the required $n = 193$ (Table 3.4.1). Thus, relatively large sample sizes are necessary to establish the negligibility of an ES. But if nothing else, this procedure at least makes explicit what it takes to say or imply from a failure to reject the null hypothesis that there is no (nontrivial) correlation or difference between A and B.

1.6 SIGNIFICANCE TESTING

Although the major thrust of this work is power analysis, a simple relationship between power and significance made it relatively simple in the computation of the power tables to provide an aid to significance testing which users of this handbook may find convenient. Generally, we can define the effect size *in the sample* (ES_s) using sample statistics in the same way as we define it for the population, and a statistically significant ES_s is one which exceeds an appropriate criterion value. For most of the power tables, these criterion values for significance of the sample ES (for the given α significance criterion and n) are provided in the second column of the power tables under the symbol for the ES for that test with subscript c (for criterion), e.g., d_c for the t test on means.

1.7 PLAN OF CHAPTERS 2-9

Each of the succeeding chapters presents a different statistical test. They are similarly organized, as follows:

Section 1. The test is introduced and its uses described.

Section 2. The ES index is described and discussed in detail.

Section 3. The characteristics of the power tables and the method of their use are described and illustrated with examples.

Section 4. The characteristics of the sample size tables and the method of their use are described and illustrated with examples.

Section 5. The use of the power tables for significance tests is described and illustrated with examples.

Statistical Power Analysis for the Behavioral Sciences

Revised Edition

Jacob Cohen

*Department of Psychology
New York University
New York, New York*



LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS
Hillsdale, New Jersey

London

Copyright © 1987 by Lawrence Erlbaum Associates, Inc.

All rights reserved. No part of this book may be reproduced in any form, by photostat, microform, retrieval system, or any other means, without the prior written permission of the publisher.

to M.

Lawrence Erlbaum Associates, Inc., Publishers
365 Broadway
Hillsdale, New Jersey 07642

Originally published 1977.

Library of Congress Cataloging in Publication Data

Cohen, Jacob, Date
Statistical power analysis for the behavioral sciences.

Bibliography: p.

Includes index.

1. Social sciences—Statistical methods.

2. Probabilities.	I. Title.	
HA29.C66 1976	300'.1'82	76-19438
ISBN 0-12-179060-6		

PRINTED IN THE UNITED STATES OF AMERICA

10 9 8 7 6 5 4 3 2 1

Contents

Preface to the Revised Edition

xi

Preface to the Original Edition

xiii

Chapter 1. The Concepts of Power Analysis

- 1.1. General Introduction
- 1.2. Significance Criterion
- 1.3. Reliability of Sample Results and Sample Size
- 1.4. The Effect Size
- 1.5. Types of Power Analysis
- 1.6. Significance Testing
- 1.7. Plan of Chapters 2-9

1
4
6
8
14
17
17



Chapter 2. The t Test for Means

- 2.1. Introduction and Use
- 2.2. The Effect Size Index: d
- 2.3. Power Tables
- 2.4. Sample Size Tables
- 2.5. The Use of the Tables for Significance Testing

19
20
27
52
66

Chapter 3. The Significance of a Product Moment r_s

- 3.1. Introduction and Use
- 3.2. The Effect Size: r

75
77

3.3. Power Tables	83
3.4. Sample Size Tables	99
3.5. The Use of the Tables for Significance Testing of r	105

9.3. F
9.4. t

Chapter 4. Differences between Correlation Coefficients

4.1. Introduction and Use	109
4.2. The Effect Size Index: q	110
4.3. Power Tables	116
4.4. Sample Size Tables	133
4.5. The Use of the Tables for Significance Testing	139

Chapt

10.1. t
10.2. t
10.3. t
10.4. t
10.5. t
10.6. t
10.7. C
10.8. F
10.9. F

Chapter 5. The Test that a Proportion is .50 and the Sign Test

5.1. Introduction and Use	145
5.2. The Effect Size Index: g	147
5.3. Power Tables	150
5.4. Sample Size Tables	166
5.5. The Use of the Tables for Significance Testing	175

Refere

Index

Chapter 6. Differences between Proportions

6.1. Introduction and Use	179
6.2. The Arcsine Transformation and the Effect Size Index: h	180
6.3. Power Tables	185
6.4. Sample Size Tables	204
6.5. The Use of the Tables for Significance Testing	209

Chapter 7. Chi-Square Tests for Goodness of Fit and Contingency Tables

7.1. Introduction and Use	215
7.2. The Effect Size index: w	216
7.3. Power Tables	227
7.4. Sample Size Tables	252

Chapter 8. F Tests on Means in the Analysis of Variance and Covariance

8.1. Introduction and Use	273
8.2. The Effect Size Index: f	274
8.3. Power Tables	288
8.4. Sample Size Tables	380
8.5. The Use of the Tables for Significance Testing	403



Chapter 9. F Tests of Variance Proportions in Multiple Regression/Correlation Analysis

9.1. Introduction and Use	407
9.2. The Effect Size Index: f^2	410